KEYNOTE

Protein structure prediction and analysis as a tool for functional genomics

Edward N Baker, Vickery L Arcus, J Shaun Lott

Centre of Molecular Biodiscovery and School of Biological Sciences, University of Auckland, Auckland, New Zealand

Abstract: Bioinformatic analyses of whole genome sequences highlight the problem of identifying the biochemical and cellular functions of the many gene products that are at present uncharacterised. Determination of their three-dimensional structures, either experimentally or by prediction, provides a powerful tool to address function, since it is at this level that biological activity is expressed. Here, we discuss the current approaches to protein structure prediction from sequence data, including the ab initio prediction of new folds, methods of fold recognition and comparative modelling based on homology. The value and limitations of such models are also explored. A major factor for the future will be the growth of the database of experimentally determined protein structures, through structural genomics projects. The prospects for this approach are also discussed, together with our experience in a pilot structural genomics project focused on proteins from *Mycobacterium tuberculosis*, the cause of tuberculosis (TB).

Keywords: protein structure, structure prediction, modelling, structural genomics, structural biology

Introduction

The explosive growth in the number of fully sequenced genomes offers an unparalleled opportunity for the understanding of organisms at the molecular level. Currently, complete genome sequences are available for more than 100 species, including key animal and plant species, and many microorganisms. At the same time, bioinformatic analyses of the gene sequences highlight the extent of our current ignorance.

The challenge of interpreting and using genomic sequence data can be illustrated by the genome sequence for *Mycobacterium tuberculosis*, the cause of tuberculosis (TB). The *M. tuberculosis* genome (Cole et al 1998) comprises about 3800 open reading frames (ORFs), each assumed to be a gene that codes for a protein product. Approximately 52% of these gene products are now annotated with a biochemical function (Camus et al 2002), with the remaining 48% being of unknown function. The latter include conserved hypotheticals (~30%), which are found in many organisms, together with a substantial population of ORFs that are found only in this or very closely related organisms. In all genomes, however, the true level of functional knowledge is probably much less, perhaps no more than 30%–40% of gene products, since most annotations have been inferred on the basis of sequence similarity to proteins from other organisms, and many are unspecific (transcription factor, oxidoreductase and so on).

Full realisation of the information encoded in genome sequences requires knowledge of the three-dimensional (3-D) structures of gene products, since it is at this level that gene function is expressed. Protein 3-D structure has traditionally provided the basis for understanding functions that have already been determined biochemically, and for applications in medicine and biotechnology such as protein engineering and structure-based drug design. Now, however, with increased throughput, it also offers a route to the discovery of function for the many gene products that are currently uncharacterised (Burley et al 1999; Mittl and Grutter 2001; Teichmann et al 2001).

Protein folds, superfamilies, families and domains

Protein structures can be classified in a hierarchical fashion, in terms of folds, superfamilies and families. Two broadly similar classifications are offered by the SCOP (Murzin et al 1995) and CATH (Orengo et al 1997) databases. The term fold describes the folding of the polypeptide chain of a protein, and may be defined in terms of how its secondary structure elements are connected (their topology) and packed

Correspondence: Edward N Baker, Centre of Molecular Biodiscovery and School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand; tel +64 9 373 7599 ext 84415; fax +64 9 373 7619; email ted.baker@auckland.ac.nz

together (Figure 1). Superfamilies comprise groups of proteins that share the same fold, and are probably related by a common evolutionary origin, but have low levels of sequence identity. Families comprise groups of proteins that have significant sequence identity and certainly share a common evolutionary origin. The term sequence family is also used to denote families within which the average pairwise sequence identity is higher than a cutoff of 30%–35%.

An important point to note is that proteins of greater than about 150 amino acid residues are usually folded into two or more structural domains. These represent autonomously folded units, typically with a hydrophobic core and a hydrophilic exterior. Since folding occurs at the level of domains, it is logical that both the description of protein folds and the prediction of protein structure should also be carried out at the level of domains. Note that structural domains do not necessarily equate to 'functional domains', and the correct assignment of structural domains within sequences is a major challenge for structure prediction (Marsden et al 2002; Galzitskaya and Melnik 2003).

A key concept of structural bioinformatics, and a reason why protein structure has potentially strong predictive power, is that of redundancy. This arises from the evolutionary nature of biology. As sequences evolve, mutations are acceptable so long as a stably folded structure

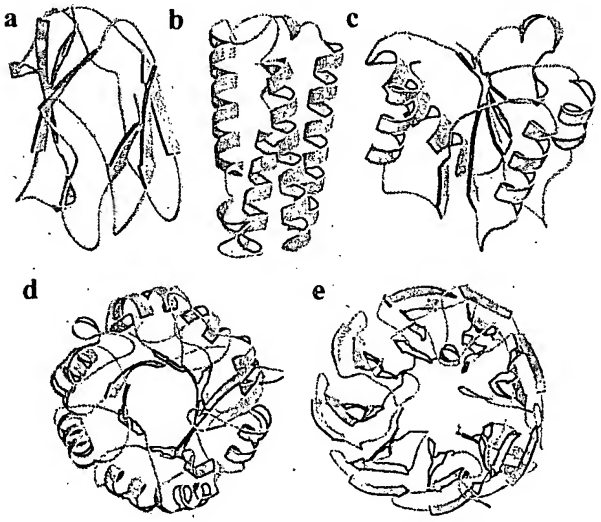

Figure 1 Examples of different protein folds. In these ribbon diagrams α-helices are shown as coils and β-strands as arrows. Folds illustrated here are **(a)** 4-helix bundle (cytochrome c'. Protein Data Bank (PDB) code 1cgn), **(b)** β-sandwich (plastocyanin, PDB code 1ag6), **(c)** open α/β domain (flavodoxin, PDB code 4fxn), **(d)** 8-stranded α/β barrel (HisF, PDB code 1h5y), and **(e)** β-propellor (alcohol dehydrogenase, PDB code 1kv9).

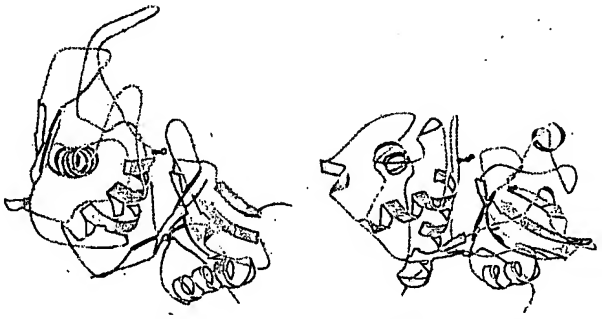

Figure 2 An example of structural conservation in distant evolutionary relatives, in the absence of significant sequence identity. On the left is SpeB, the cysteine protease from *Streptococcus pyogenes* (PDB code 1dki) and on the right actinidin, a similar enzyme from kiwifruit (PDB code 2act). Note that although SpeB has additional embellishments both proteins have the same basic fold and a catalytic cysteine residue in the same position. The two proteins share only 8% sequence identity.

is maintained, and although we do not fully understand the sequence 'codes' that lead to a given protein fold, there is evidently considerable redundancy. Residues in some parts of a protein structure can change freely, and others may be able to change so long as a certain character is maintained. The result is that proteins can diverge beyond significant sequence similarity but still retain the 3-D fold of their ancestor (see for example Figure 2). Therefore, although there are many millions of unique sequences, the number of protein families is more limited. Various calculations have attempted to extrapolate from the current database to estimate the numbers of folds, families and superfamilies. These suggest that the number of unique folds is as little as 1000–4000 (Chothia 1992; Govindarajan et al 1999; Wolf et al 2000). The number of families is larger, and may be between about 5000 and 50 000 depending on what level of sequence identity is used to define a family; the figure of 50 000 may be relevant at a level of 30% sequence identity (Coulson and Moult 2002).

Structure prediction and modelling

The central problem in predicting the 3-D structure of a protein from its amino acid sequence alone is that it is still not understood what 'code' in its sequence directs a given polypeptide towards a particular fold so rapidly and reproducibly. This is the so-called protein folding problem. Early events include a 'hydrophobic collapse', in which hydrophobic side chains cluster together in a molten globule core, and the formation of secondary structures (Baldwin 1989), and it has also become apparent that local interactions play a significant role (Chan 1998).

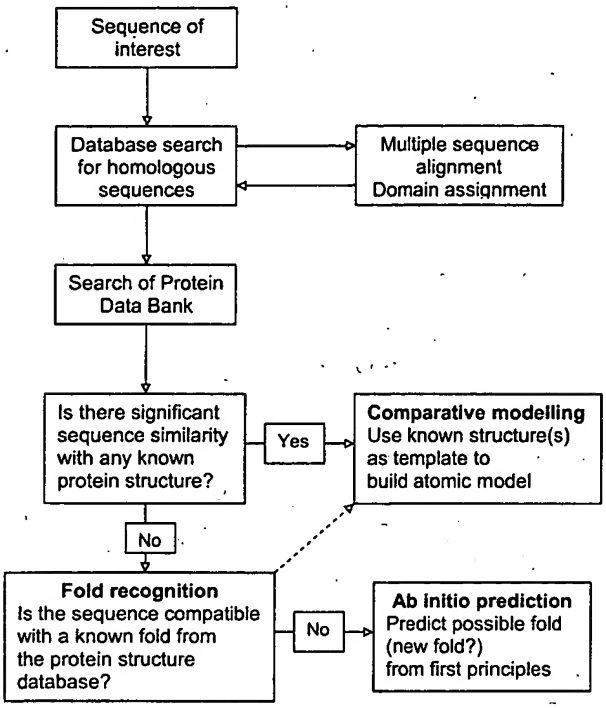

Figure 3 Flow chart showing the various steps and options for predicting the three-dimensional structure of a protein.

Prediction methods can be loosely categorised as either ab initio or knowledge-based approaches, the latter making use of the growing protein structure database. Ab initio methods provide the only route to structure prediction for new folds (ie folds that are not yet found in the experimental database). Knowledge-based methods, on the other hand, depend on recognising that the sequence in question either is compatible with an already known fold (fold recognition) or is similar enough to that of some known protein structure that an atomic model can be built (comparative modelling). These options are set out in Figure 3, and a number of relevant databases and websites are listed in Table 1.

Ab initio methods

Ab initio methods attempt to generate structural models solely on the basis of the principles of physics and chemistry. Despite a variety of innovative approaches over the years these have met with little success. Recently, however, some very promising results have been achieved with a program called Rosetta (Simons et al 1997). In this method, the importance of local interactions is recognised. The polypeptide in question is divided into short segments (typically 3 and 9 residues) that are allowed to continually sample possible local conformations until combinations are found that are of low energy and (a) bury hydrophobic side chains, and (b) result in the pairing of any β-strands. The method does include an element of knowledge-based modelling because the local conformations that are sampled by the short peptides are taken from conformations found already in experimentally determined structures. The method has, however, shown impressive success in predicting new folds in the critical assessment of methods of protein structure prediction (CASP) assessment exercises (Bonneau et al 2001).

Fold recognition methods

Fold recognition is based on the principle that every distinct protein fold has its own pattern of features (buried hydrophobic residues, secondary structures etc) with which its amino acid sequence must be compatible. Thus, it should be possible to assess the compatibility of a given sequence with each of the various known folds.

An early approach to fold recognition made use of secondary structure prediction methods, such as those in the classic Chou–Fasman approach (Chou and Fasman 1978) in which residues in a sequence were assigned empirical preferences for α-helix, β-strand or 'random coil' conformations, which were then used to assign secondary structure elements. Although success rates did not exceed ~60% for single sequences, multiple sequence alignments increased the reliability of secondary structure prediction. In combination with hydropathy and flexibility profiles this

Table 1 Useful web addresses[a]

Protein Data Bank (database of 3-D structures)	
	http://www.rcsb.org/pdb/
Classifications of protein structure	
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/
HOMSTRAD	http://www-cryst.bioc.cam.ac.uk/data/align/
Some structure prediction servers	
3D-PSSM	http://www.sbg.bio.ic.ac.uk/3dpssm/
GenTHREADER	http://bioinf.cs.ucl.ac.uk/psipred/
FUGUE	http://www-cryst.bioc.cam.ac.uk/fugue/
Assessment of structure prediction methods	
CASP	http://PredictionCenter.llnl.gov/
LiveBench	http://bioinfo.pl/LiveBench/
EVA	http://cubic.bioc.columbia.edu/eva/
Structural genomics initiatives	
NIH protein structure initiative	http://www.structuralgenomics.org/
TB consortium	http://www.doe-mbi.ucla.edu/TB/
SPinE (European consortium)	http://www.spineurope.org/

[a] All websites accessed 7 Jul 2003.

approach successfully predicted an 8-stranded α/β barrel structure for the α-subunit of tryptophan synthase (Crawford et al 1987).

A more systematic approach to fold recognition is provided by so-called 'threading' methods. In one form of this (Bowie et al 1991), a 3-D to 1-D profile (a structure profile) is created for each unique fold in the Protein Data Bank. Every amino acid position in the polypeptide is coded with a description of its environment, in terms of its solvent accessibility, the polarity of its environment and its secondary structure location. A new sequence of interest is then threaded on to the structure profile of a given fold, and its compatibility assessed with a scoring function. In this way it can be tested against all the known protein folds. An important feature of this approach is that it is not dependent on sequence alignment and in principle can therefore find proteins that have the same fold but no detectable sequence similarity.

A number of variations on this threading approach exist. Instead of the environment description of the 3-D to 1-D profile, energy potentials can be used (Jones et al 1992). The strength of the predictions is enhanced by multiple alignments of homologous sequences, as in GenTHREADER (Jones 1999), and will presumably increase as the sequence database grows. Several algorithms combine secondary structure and accessibility predictions with threading. In this approach, which is used by 3D-PSSM (Kelley at al 2000), a profile is generated for the sequence in question (and its homologues) and is threaded on to the corresponding profiles of all known folds from the present structural database. More sophisticated methods add further structural attributes. For example, the program FUGUE (Shi et al 2001) incorporates environment-specific amino acid substitution tables, structure-dependent gap penalties, and information from both multiple sequence alignments and multiple structure alignments. The philosophy here is that since the goal is structure prediction, and since structure is much more strongly conserved than sequence during evolution, the incorporation of structural knowledge should lead to more powerful predictions.

Comparative modelling

Comparative modelling, also known as homology modelling, takes advantage of the structural similarities within protein families. The assumption is that all the members of a protein family are related by divergent evolution from a common ancestor (ie are homologous) and must therefore share the same basic fold. Thus, if a new protein sequence is found (by sequence alignment) to belong to a recognisable protein family, and 3-D structures are already available for one or more members of that family, an atomic model can be built by comparison with those structures.

Basic approaches to comparative modelling were set out by Greer (1981), and by Sali and Blundell (1993), and an up-to-date review is given by Contreras-Moreira et al (2003). The first step is to generate a template structure. If structures are available for two or more family members, either a single one is chosen, or an average of all available structures. If multiple structures are available, superposition will enable them to be described in terms of structurally conserved regions (SCRs), that align very closely, and variable regions (VRs). Typically, the SCRs include core secondary structural elements and the active site, and VRs include the connecting loops between secondary structure elements. The sequence in question is then aligned with the template using, for example, PSI-BLAST or sequence signatures that are characteristic of the SCRs. Construction of an atomic model is achieved by fitting the sequence in question to the template. This is done either for the whole polypeptide simultaneously (Sali and Blundell 1993) or sequentially (Greer 1981). The latter approach involves modelling first the SCRs, which can be built with a high degree of confidence, then the VRs and then the side chains. There are many methods that specialise in the modelling of loops (Fiser et al 2000) and side chains (Bower et al 1997) within the restrained environment provided by the rest of the structure. Finally, the model can be refined, for example by simulated annealing, and assessed.

Assessment of prediction methods

A major contribution to the protein structure prediction field has been made by the ongoing CASP experiment. Designed by John Moult and colleagues, and carried out every 2 years, this experiment submits structure prediction methods to a blind test (Moult et al 2001). Protein structures that have been determined experimentally are withheld from publication while structure prediction groups are invited to tender their predictions; the resulting predictions are then assessed against the experimental structures. Targets have traditionally been divided into three different categories, ab initio prediction, fold recognition and comparative modelling. In the CASP4 and CASP5 competitions, the ab initio category has been renamed as new fold recognition, prompted by the development of methods such as Rosetta. Results are presented in special issues of the journal

Proteins, the latest (CASP4) in 2001 (Moult et al 2001), and can also be found online (http://PredictionCenter.llnl.gov/).

An interesting assessment of the success of automated structure prediction servers is provided by the LiveBench experiment (http://bioinfo.pl/LiveBench/), which automatically submits every new structure deposited in the Protein Data Bank to the various fold recognition servers and maintains a ranking order. A similar experiment, EVA (http://cubic.bioc.columbia.edu/eva/), focuses on secondary structure prediction.

Current state of protein structure prediction

From the results of the CASP exercises a number of comments can be made on the current state of protein structure prediction methods (see also Schonbrun et al 2002).

- In the ab initio (new fold) structure prediction category, the most exciting development has been the success of the Rosetta method, developed by David Baker and colleagues. This approach has produced some excellent fold predictions (Bonneau et al 2001), although the problem remains of picking a correct solution from incorrect ones. Structural models from this method may allow functional predictions to be made, but are essentially of low resolution at this moment.
- Fold recognition methods work better for distant homologues, than for analogues (unrelated proteins with the same fold) because of the extra information that comes from multiple sequence alignments. The main problems are in pattern degeneracy (eg the patterns of hydrophobic residues can look quite similar between different folds), and in the considerable divergence of structure that occurs between very distant homologues (de la Cruz and Thornton 1999). This has the effect that although the fold may be correct, the detailed alignment of the sequence with the fold seldom is. Again this gives essentially low-resolution information, albeit with considerable value for predicting function, catalytic residues, binding sites and so on.
- The quality and usefulness of comparative models (homology models) depends critically on the level of sequence identity (Baker and Sali 2001). Above 50% sequence identity the models are excellent, comparable with experimental nuclear magnetic resonance (NMR) structures, or medium resolution crystal structures, and can be used for analysing catalytic mechanisms, docking and improving ligands etc. In the range 30%–50% identity, up to 90% of the polypeptide conformation tends to be modelled well, with an accuracy of ~1.5 Å, suitable for many functional purposes (eg site-directed mutagenesis). Below 30% sequence identity, however, there are increasing alignment errors and this is really the cutoff for effective atomic-level modelling.
- Automated servers are now producing impressive results but still cannot match the power of experienced human intervention. A case in point from the CASP4 experiment was Alexei Murzin's recognition in one target that a helix and strand in the template had to be replaced in the target by a loop and a strand going in the opposite direction (Murzin and Bateman 2001).
- A major problem is that no structure refinement method yet exists that can correct errors of modelling effectively, particularly alignment errors. It has been a frequent observation that refinement of homology models usually makes them worse (ie takes them further away from the true structure) and the original template is often closer (Tramontano et al 2001). This is an area that certainly needs improvement.
- All modelling methods share some common drawbacks. Bound ligands, metal ions or water molecules will not be predicted. Structure prediction necessarily is performed at the level of domains, and we have no effective way, as yet, of predicting the association of domains in multidomain proteins. This is important because active sites are often at domain interfaces. And finally there remains the thorny issue of assessing the validity of any model. Caveat emptor!

Experimental approaches: the promise of structural genomics

Two methods are available for the experimental determination of protein 3-D structure: x-ray crystallography and NMR. Both have seen major technological improvements in recent years that have dramatically increased the speed of structure determination. In the case of x-ray crystallography, more systematic crystallisation methods, crystal freezing, automated phasing and model building methods, and in particular, the use of synchrotrons and the MAD phasing method, have revolutionised the field. In the case of NMR, higher-field spectrometers and innovative new methods have steadily increased the practical size limit. These advances raise the possibility of high-throughput structure determination as a real possibility for addressing

functional genomics (Kim 1998; Sali 1998; Terwilliger et al 1998; Burley et al 1999; Mittl and Grutter 2001).

The concept of structural genomics is based on the idea that since function depends on protein 3-D structure, and structure is conserved much more strongly than sequence, it should be possible to use 3-D structure determination to help discover function. There would also be the bonus that the resulting protein structures would (a) provide the basis for new drug discovery or protein engineering, and (b) expand the structural database so as to greatly increase the ability to model other proteins.

These ideas have prompted major investments in structural genomics in a number of countries, for example the NIH-sponsored Protein Structure Initiative in the United States, the Japanese Protein3000 project, the European Structural Proteomics in Europe (SPinE) collaboration, and other smaller projects in Germany, Canada, UK, France, Finland, Israel and other countries. In addition, several structural genomics companies (Syrrx and Structural Genomix in the United States, Affinium in Canada) have been formed. The specific aims of structural genomics include the development of high-throughput structure determination methods, the prediction of function from structure, the determination of experimental structures for at least one member of every protein family (such that all other family members will be within modelling distance), and the identification and characterisation of potential drug targets. One estimate is that 16000 protein structures, carefully chosen, would achieve the goal of having at least one representative structure for every family (Vitkup et al 2001). This would largely bypass the protein folding problem, as comparative modelling could then be used for homologous sequences.

The experimental challenge of structural genomics is considerable, since large-scale structure determination requires an inherently more complex, and less automatable, set of steps than DNA sequencing. These include gene cloning, protein expression, purification and crystallisation, and rapid structure determination by crystallography or NMR. Progress in our own laboratory in the past 3 years, as part of the TB Structural Genomics Consortium (see below) illustrates the bottlenecks in such a project and the diminishing returns at each step. With a team of 3 postdoctoral researchers and several students, we have cloned the genes for 86 proteins, obtained expression for 78 of them (but only 38 in soluble form), crystallised 20, and determined 8 structures to date. The bottlenecks are in obtaining soluble expression and in crystallisation. (Note that in this and other structural genomics enterprises, membrane proteins are deliberately excluded as too hard: they are a challenge for the future.)

Our own involvement in structural genomics is as partners in the TB Structural Genomics Consortium (TBSGC) (http://www.doe-mbi.ucla.edu/TB/). Centred in the United States as one of nine structural genomics consortia under the umbrella of the NIH Protein Structure Initiative (http://www.structuralgenomics.org/), the TBSGC focuses on *Mycobacterium tuberculosis* with the goal of large-scale determination of protein structures that will aid in the development of new TB drugs and in understanding TB biology. Although NIH funding is limited to the US participants (Los Alamos National Laboratory, UCLA, UC Berkeley, Texas A&M, Albert Einstein College of Medicine), laboratories from 10 countries participate and share the new technologies that are developed. They have access to central bioinformatic resources and facilities for protein expression and crystallisation, gene knockouts and synchrotron data collection.

What can be learned about function from structural genomics?

Protein structures offer a range of possibilities for the discovery of function (Eisenstein et al 2000; Teichmann et al 2001; Zhang and Kim 2003). Careful examination of surface clefts or cavities, or mapping electrostatic charge, may indicate a possible active site or binding site. Mapping conserved residues on to the fold may likewise identify a key site or perhaps a characteristic catalytic motif. For protein structures determined by crystallography there is the possibility that a bound cofactor or ligand may be found associated with the protein; even adventitiously bound solvent molecules may be of value here. In a few cases very powerful insights into function are gained by the discovery of unexpected structural homologies that were not apparent at the sequence level. Examples from our own work illustrate several of these possibilities.

Pa_989. This protein, from the thermophilic organism *Pyrobaculum aerophilum*, is homologous with HisF, an enzyme from the histidine biosynthetic pathway. The structure, determined at 2.0 Å resolution, revealed an 8-stranded α/β barrel fold that immediately identified the probable location of the active site at one end of the barrel.

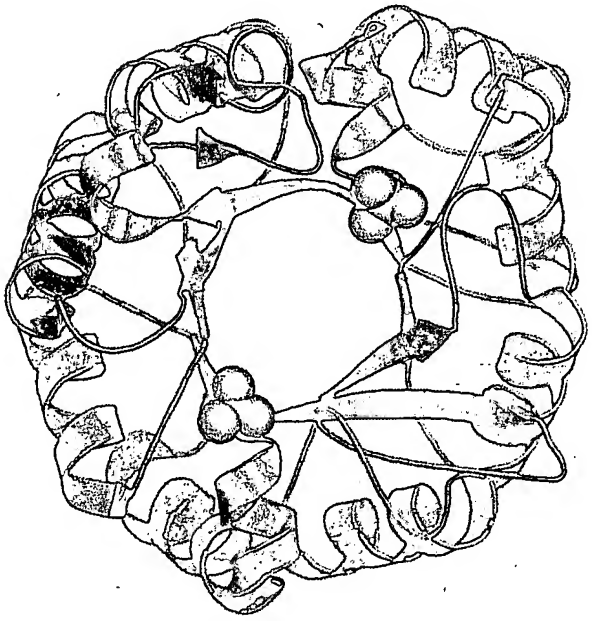

Figure 4 Functional clues from crystal structure analysis. In the crystal structure of HisF (Banfield et al 2001), two adventitiously-bound phosphate ions (from the buffer solution used) mark out the binding site for the substrate, a biphosphorylated molecule.

More revealingly, two phosphate ions, adventitiously bound from the buffer, marked out the precise location where the substrate (a bisphosphorylated molecule) would bind (Figure 4) (Banfield et al 2001).

Pa_2307. This *P. aerophilum* protein, annotated as a conserved hypothetical domain found in many species including *M. tuberculosis*, was determined at 1.6 Å resolution. The fold gives few clues to function, as it has been seen just once before, in a domain of unknown function from an RNA polymerase. The structure is found, however, to include a phosphorylated histidine residue that presumably marks the active site and suggests a role in phosphate transfer.

Rv3853. From *M. tuberculosis*, this protein was annotated in the genome sequence as MenG, a SAM-dependent methyltransferase from the menaquinone biosynthetic pathway. The trimeric structure proved, however, to be unlike that of any other methyltransferase and strongly suggests that the functional annotation is incorrect (Johnston et al 2003). The true function remains unclear, but a groove at the subunit boundary flanked by a number of conserved residues and harbouring several small molecule ligands, suggests a binding site for an extended ligand, perhaps a polypeptide or nucleic acid strand.

SpeB. This protein from *Streptococcus pyogenes* was known to be a cysteine protease implicated in the 'flesh-eating' disease, necrotising fasciitis. Its amino sequence showed no detectable similarity with any other in the sequence database. Unexpectedly, however, the crystal structure (Kagawa et al 2000) showed that SpeB clearly belongs to the papain superfamily, with the same fold despite a sequence identity of only 8% (Figure 2). This is a clear example where even if SpeB had not been known to have cysteine protease activity, one look at the structure would have been sufficient to reveal its function.

Conclusions

Several factors suggest that we can expect dramatic increases in the availability of protein 3-D structural information in the next few years. First, protein structure prediction methods are steadily improving. Second, both the sequence and structure databases are expanding rapidly, the former through genome sequencing projects and the latter through the impact of structural genomics. Not only will these developments enhance the power of prediction methods, but there is also a realistic prospect that most sequences, at least of non-membrane proteins, will be within modelling distance of an experimentally determined structure. A key area for the bioinformatics of the future, then, will be to learn to extract functional information from protein structures rather than protein sequences.

Acknowledgements

This work is supported by CoRE funding to the Centre of Molecular Biodiscovery, and by grants from the Health Research Council of New Zealand, the Marsden Fund, and the New Economy Research Fund.

References

Baker D, Sali A. 2001. Protein structure prediction and structural genomics. *Science*, 294:93–6.

Baldwin RL. 1989. How does protein folding get started? *Trends Biochem Sci*, 14:291–4.

Banfield MJ, Lott JS, Arcus VL, McCarthy AA, Baker EN. 2001. Structure of HisF, a histidine biosynthetic protein from *Pyrobaculum aerophilum*. *Acta Crystallogr D Biol Crystallogr*, 57:1518–25.

Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. 2001. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins*, 45(Suppl 5):119–26.

Bower MJ, Cohen FE, Dunbrack RL Jr. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modelling tool. *J Mol Biol*, 267:1268–82.

Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–9.

Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. 1999. Structural genomics: beyond the Human Genome Project. *Nat Genet*, 23:151–7.

Camus J-C, Pryor MJ, Medigue C, Cole ST. 2002. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology*, 148:2967–73.

Chan HS. 1998. Matching speed and locality. *Nature*, 392:761–3.

Chothia C. 1992. Proteins – 1000 families for the molecular biologist. *Nature*, 357:543–4.

Chou PY, Fasman GD. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol*, 47:54–148.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393:537–44.

Contreras-Moreira B, Fitzjohn PW, Bates PA. 2003. Comparative modelling: an essential methodology for protein structure prediction in the post-genomic era. *Appl Bioinform*, 1:177–90.

Coulson AFW, Moult J. 2002. A unifold, mesofold, and superfold model of protein fold use. *Proteins*, 46:61–71.

Crawford IP, Niermann T, Kirschner K. 1987. Prediction of secondary structure by evolutionary comparison: application to the α subunit of tryptophan synthase. *Proteins*, 2:118–29.

de la Cruz X, Thornton JM. 1999. Factors limiting the performance of prediction-based fold recognition methods. *Protein Sci*, 8:750–9.

Eisenstein E, Gilliland G, Herzberg O, Moult J, Orban J, Poljak RJ, Banerjei L, Richardson D, Howard AJ. 2000. Biological function made crystal clear – annotation of hypothetical proteins via structural genomics. *Curr Opin Biotechnol*, 11:25–30.

Fiser A, Do RKG, Sali A. 2000. Modeling of loops in protein structures. *Protein Sci*, 9:1753–73.

Galzitskaya OV, Melnik BS. 2003. Prediction of protein domain boundaries from sequence alone. *Protein Sci*, 12:696–701.

Govindarajan S, Recabarren R, Goldstein RA. 1999. Estimating the total number of protein folds. *Proteins*, 35:408–14.

Greer J. 1981. Comparative model-building of the mammalian serine proteases. *J Mol Biol*, 153:1027–42.

Johnston JM, Arcus VL, Morton CJ, Parker MJ, Baker, EN. 2003. The crystal structure of a putative methyltransferase from *Mycobacterium tuberculosis*: misannotation of a genome clarified by protein structural analysis. *J Bacteriol*, 185:4057–65.

Jones DT. 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*, 287:797–815.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature*, 358:86–9.

Kagawa T, Cooney JC, Baker HM, McSweeney S, Liu M, Gubba S, Musser JM, Baker EN. 2000. Crystal structure of the zymogen form of the group A *Streptococcus* virulence factor SpeB: an integrin-binding cysteine protease. *Proc Natl Acad Sci USA*, 97:2235–40.

Kelley LA, MacCallum RM, Sternberg MJE. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol*, 299:499–520.

Kim SH. 1998. Shining a light on structural genomics. *Nat Struct Biol*, 5:643–5.

Marsden RL, McGuffin LJ, Jones DT. 2002. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci*, 11:2814–24.

Mittl PRE, Grutter MG. 2001. Structural genomics: opportunities and challenges. *Curr Opin Chem Biol*, 5:402–8.

Moult J, Fidelis K, Zemla A, Hubbard T. 2001. Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins*, 45(Suppl 5):2–7.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247:536–40.

Murzin AG, Bateman A. 2001. CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. *Proteins*, 45(Suppl 5):76–85.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH – a hierarchic classification of protein domain structures. *Structure*, 5:1093–108.

Sali A. 1998. 100 000 protein structures for the biologist. *Nat Struct Biol*, 5:1929–32.

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 269:423–39.

Schonbrun J, Wedemeyer WJ, Baker D. 2002. Protein structure prediction in 2002. *Curr Opin Struct Biol*, 12:348–54.

Shi J, Blundell TL, Mizuguchi K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, 310:243–57.

Simons KT, Kooperberg C, Huang E, Baker D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, 268:209–25.

Teichmann SA, Murzin AG, Chothia C. 2001. Determination of protein function, evolution and interactions by structural genomics. *Curr Opin Struct Biol*, 11:354–63.

Terwilliger TC, Waldo G, Peat TS, Newman JA, Chu K, Berendzen J. 1998. Class-directed structure determination: foundation for a protein structure initiative. *Protein Sci*, 7:1851–6.

Tramontano A, Leplae R, Morea V. 2001. Analysis and assessment of comparative modeling predictions in CASP4. *Proteins*, 45(Suppl 5):22–38.

Vitkup D, Melamud E, Moult J, Sander C. 2001. Completeness in structural genomics. *Nat Struct Biol*, 8:559–66.

Wolf YI, Grishin NV, Koonin EV. 2000. Estimating the total number of protein folds and families from complete genomic data. *J Mol Biol*, 299:897–905.

Zhang C, Kim S-H. 2003. Overview of structural genomics: from structure to function. *Curr Opin Chem Biol*, 7:1–5.